Vowel Classification for Computer-Based Visual Feedback for Speech Training for the Hearing Impaired

Stephen A. Zahorian, A. Matthew Zimmer, and Fansheng Meng

Department of Electrical and Computer Engineering, Old Dominion University
Norfolk, Virginia 23529, USA

ABSTRACT

A visual speech training aid for persons with hearing impairments has been developed using a Windows-based multimedia computer. The training aid provides real time visual feedback as to the quality of pronunciation for 10 steady-state American English monopthong vowels (/aa/, /iy/, /uw/, /ae/, /er/, /ih/, /eh/, /ao/, /ah/, and /uh/). This training aid is thus referred to as a Vowel Articulation Training Aid (VATA). Neural network (NN) classifiers are used to classify vowels and then provide real time feedback for several displays: a 10-category "vowel bargraph" which provides "discrete" feedback, an "ellipse display" which provides continuous feedback over a 2-D space similar to a formant1-formant2 space, and three game displays (a form of "tetrus," controlled by one vowel, a "chicken crossing the road," controlled by two vowels, and pacman, controlled by four vowels.) Continuous feedback such as this is desirable for speech training to help improve articulation. In this paper we describe the overall speech training system, discuss some algorithmic refinements to the vowel classifier, and report some experiments related to the development of a database used for "training" the display.

1. BACKGROUND

Displays Available

The system has two main displays. One is a bargraph display, which gives feedback about how well speech utterances match discrete vowel categories. The other is an "ellipse" display, which provides a more continuous feedback about vowel pronunciation. The system is designed to provide feedback for /ah/, /ee/, /ue/, /ae/, /ur/, /ih/, /eh/, /aw/, /uh/, and /oo/, which correspond to the vowel sounds found in the words "cot," "beet," "boot," "bag," "bird," "pig," "bed," "dog," "cup," and "book" respectively. The labels for this system (/ah/, etc.) were assigned by the ODU speech lab, and correspond to the ARPABET labels /aa/, /iy/, /uw/, /ae/, /er/, /ih/, /eh/, /ao/, /ah/, and /uh/ commonly used in speech processing literature.

The bargraph display (Figure 1) resembles a histogram, with one bar for each vowel sound of interest. The height of the vowel's bar varies in proportion to the accuracy of the speaker's pronunciation of that vowel. Correct pronunciation yields one steady, clearly defined bar, while the rest assume zero or small values. Incorrectly pronounced sounds may produce displays showing two or more partially activated category bars, no activated bars, or rapid fluctuations between bars.

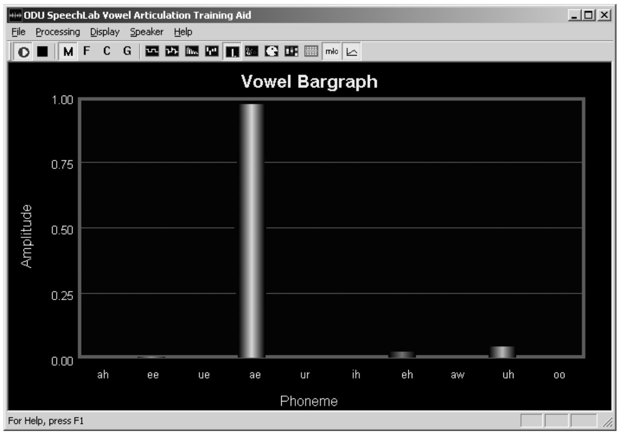

Figure 1. Bargraph display showing response for correct pronunciation of /ae/

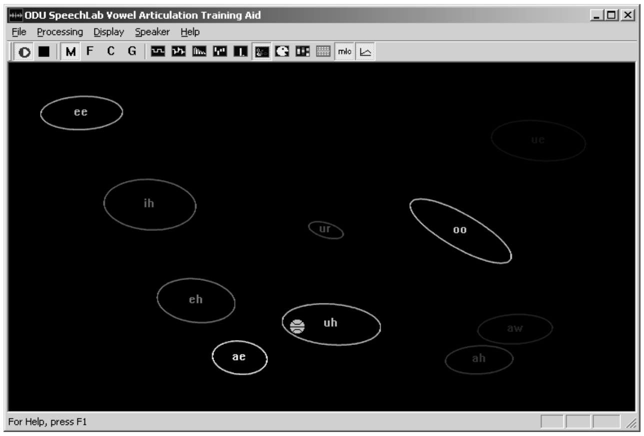
ODU SpeechLab Vowel Articulation Training Aid
File Processing Display Speaker Help
M F C G
ee
ue
ih
ur
oo
eh
uh
aw
ae
ah
For Help, press F1

Figure 2. Ellipse Display showing response for correct pronunciation of /uh/

The ellipse display (Figure 2) divides the screen into several elliptical regions, similar to a F1/F2 type display. Unlike the F1/F2 display, this 'ellipse' display bases its output on a neural network, which has been trained to transform Discrete Cosine Transform Coefficients (DCTCs) of the log magnitude spectrum to specified target positions in two-dimensional space. Correct pronunciation of a particular sound places a basketball icon within the corresponding ellipse and causes the icon's color to match that of the ellipse. Incorrect or unclear pronunciation results in the ball icon 'wandering' about the screen or coming to rest in an area not enclosed by an ellipse. By observing the continuous motion of the ball, a speaker hopefully can learn to adjust his or her pronunciation in order to produce the desired vowel sound.

In addition to the bargraph and ellipse display, three game displays have been developed. One game is a simplified version of "tetrus," for which one vowel sound is used to control the orientation of the colors in a falling bar in order to score. In another game, a chicken, controlled by two vowel sounds, attempts to navigate a busy highway without being hit by a vehicle. The third game is pacman, which uses four vowel sounds to control the four directions of movement of the game icon. Figure 3 depicts the pacman game.

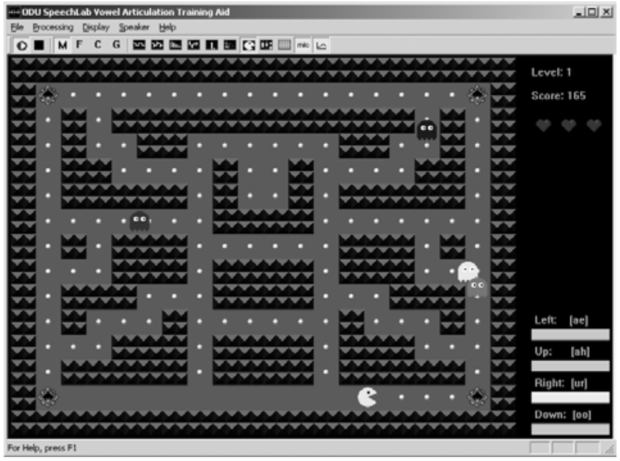

Figure 3 Pacman game for vowel training

A speaker group selection option with "CHILD," "FEMALE" and "MALE" settings allows all displays to be fine-tuned for better classification of sounds produced by child, adult female, or adult male speakers respectively. A fourth speaker group option, "GENERAL" uses a classifier based on all speakers.

2. PROCESSING STEPS

A block diagram for the VATA system is shown in Figure 4. The operating system interacts with the sound card to acquire a section of data (a "segment") from the from the continuous audio data stream. Since driver software communicates with the operating system services and not directly the hardware, the VATA system is able to work with most commonly available Windows-compatible sound cards.

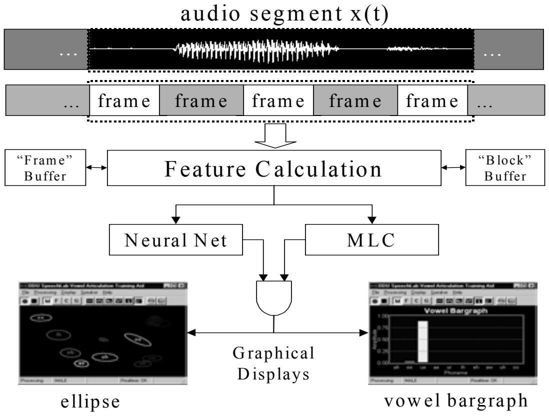

Figure 4. VATA Block Diagram

Following division of the signal into frames, signal processing is performed as shown in figure 4. Each frame is first passed through a high-frequency pre-emphasis filter, typically centered at 3.2kHz. Next a Fast-Fourier Transform (typically 512 points) is performed on each frame and the logarithm of the magnitude of the resulting coefficients is taken. This "log-magnitude spectrum" is then averaged over several (usually about 5 to 10) frames, and a Discrete Cosine Transform (DCT) expansion using is performed to yield the Discrete Cosine Transform ("Cepstral") coefficients, which are considered as the "features" of the signal. These DCTC terms, similar to cepstral coefficients, also make use of frequency warping to substantially increase vowel classification rates [2].

The first several (typically 12) cepstral coefficients (DCTCs) are then normalized (zero mean and standard deviation of ± .2, or typically a range of ±1) and passed to a neural network (NN) classifier. The neural networks used in VATA have one input node per feature, 15 to 25 nodes in the hidden layer, and ten output nodes (one per vowel class). The NN's are trained using error backpropagation for (typically) 250,000 iterations.

3. THE GAUSSIAN BAYESIAN CLASSIFIER

One of the major problems with a neural network approach for any type of pattern classification is that "out-of-category" exemplars are also classified, according to whichever trained class is "closest" to the exemplar. During testing of the VATA system, it became apparent that sounds which were not "correct" examples of any of the 10 steady-state vowels for which the system was designed caused "false" correct indications. This has the unfortunate consequence of producing feedback corresponding to a "correct" pronunciation when in fact no valid vowel sound has been uttered. For example, ambient room noise could often trigger the display for /ee/. This unintended behavior could, for a user relying solely on VATA for pronunciation feedback, cause confusion and frustration.

To reduce the number of "false correct" displays, a modified

Euclidean-Distance measure, has been incorporated with the neural network classifier as a secondary verification. The general Gaussian distance equation requires the use of the mean vector and covariance matrix for each class (computed from a database of training data) and is given by:

$$D_i(\mathbf{f}) = (\mathbf{f} - \bar{\mathbf{f}}_i)^T \mathbf{R}_i^{-1} (\mathbf{f} - \bar{\mathbf{f}}_i) + \ln|P_i| \tag{1}$$

This general equation has many parameters (and thus requires lots of training data) and is also computationally complex. For the case of the speech display, the features are approximately uncorrelated, and hence its inverse is approximately diagonal. Also the (apriori probability) term remains constant and can be ignored. Taking advantage of these two considerations, Equation 1 reduces to the much simpler form shown in Equation 2, which we refer to as the weighted modified Euclidean distance:

$$D_i(\mathbf{f}) = \sqrt{\sum_{j=1}^{m} w_j \left(\frac{\mathrm{f}_j - \bar{\mathrm{f}}_{ij}}{\sigma_{ij}} \right)^2} \tag{2}$$

These weights, $w_,$ are important to use if the "information" contained in the underlying features is not proportional to the feature variances. For the case of the DCTC features for vowel recognition, our previous work has shown that the DCTCs do not uniformly contribute to vowel recognition [2]. Based on this earlier work, relative weights of (.82, 1.65, 2.47, 2.47, 2.06, 1.65, 1.24, .83, .41, .41, .41, .21, .21, .08, .08) were used. The actual weights were the relative weights given, but normalized such that the sum of the weights was 1.0. We refer to classification based on this distance as Maximum Likelihood Classification (MLC), since maximum likelihood methods based on Gaussian assumptions are used to obtain the parameters in the classifier.

To provide verification that the vowels display is producing accurate feedback, the MLC calculates the distance of the average features for the NN's choice from the features of the token under evaluation by the NN. If the feature distance is within the threshold criterion,

$$D_i(\mathbf{f}) < \alpha\sqrt{m} \tag{3}$$

Where m is the number of features (typically 10 to 15 for VATA), and α is an arbitrary scale factor used for performance tuning, the neural network's decision is accepted, otherwise it is discarded. If is too small, the MLC will reject many correct tokens; if is too large, the out-of-category tokens will still not be rejected. Tests (described below) have shown that with the properly determined threshold (α =1.2), the MLC does reject unwanted sounds, and makes very few false rejections.

To experimentally verify the operation of the MLC method described above, experiments were done as follows: Features were computed using a database of 10 vowel sounds obtained from three speaker panels, with each speaker producing each vowel sound three times. Vowels were pronounced as "isolated" words, in response to a computer prompt. The central 200 ms interval section of each vowel was used for processing. A two- layer (one hidden layer) neural network was trained as a classifier, using nine vowels (those listed above except for /ur/). The means and standard deviations for the features for these nine vowels were also computed, on a vowel by vowel basis, since these are the features needed for the MLC. The neural network/MLC classifier system was then evaluated using test data from 24 different speakers, and using data for 9 vowels, consisting of all the training vowels, except with /ur/ substituted for /oo/. Thus, ideally the classifier should have correctly recognized 8 vowels, but rejected /ur/ as being out of category.

The experiment was repeated separately for male speakers, female speakers, and male/female speakers combined. 50 training speakers were used for the male speaker case, 50 for the female speaker case, and 80 speakers (40 male, 40 female) for the combined case. There were 24 test speakers for the male case, 24 for the female case, and 48 test speakers for the combined case. In each experiment the false acceptance rate, and false rejection rates were obtained as a function of the parameter α above. False acceptances were considered as instances of the /ur/ being accepted as any one of the vowels. False rejections were considered as instances of any vowels classified correctly by the neural network, but incorrectly rejected by the MLC.

Typical results (for the female speaker case) are shown in figure 5, as false acceptance and false rejections as a function of α. As expected for low values of α, the false rejection rate is very high. For high values of α, the false acceptance is high. Similar results were obtained for the male speakers and combined speaker case. Values of approximately 1.5 insure that most out of category tokens are rejected, but that very few in category tokens are rejected.

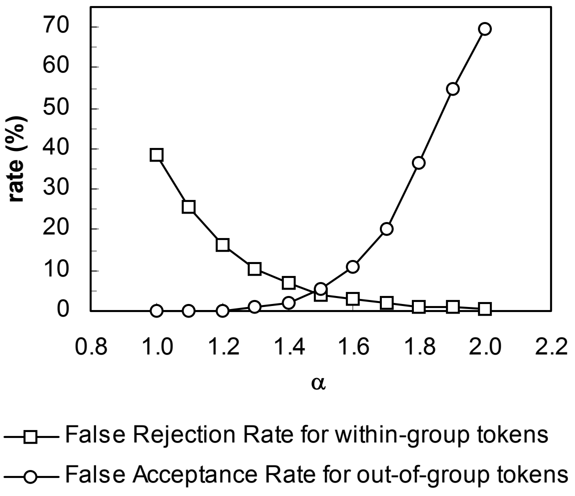

Figure 5. False acceptance/rejection rates as a function of decision threshold value α, female speaker case.

4. DATABASE SIZE AND VOWEL CLASSIFICATION ACCURACY

One of the major issues which has a big effect on the accuracy of a neural network classifier, such as the one used for the basic vowel classifier used in VATA, is the amount of training data available to train the classifier. In general, although the accuracy on training data decreases as more training data is used, classifier accuracy on testing data improves as the size of the training database increases. The size of the database needed to achieve "best"performance generally depends on the number of features used, the classifier complexity (or number of weights in a neural network classifier), and the general properties of the features.

In order to experimentally investigate database size issues in the context of the VATA system, we have recorded vowel sounds from over 400 speakers---approximately evenly divided between adult men, adult women, and children. We then conducted a series of classification tests with the number of training speakers varied from 2 to 100, the number of hidden nodes in an NN varied from 5 to 100, and the number of test speakers fixed at 25. Results of one such experiment are shown in Figure 6, for the case of female speakers. As expected, training results decrease as more training speakers are added, and classification accuracy improves for the test speakers. Similar results were obtained for the adult male speakers and female/male speakers combined. We generally found that test performance changed very little if at least 60 training speakers were used. Additionally, it was observed that performance changes very little as the number of hidden nodes changes from 25 to 100. Since the required computations are much less with 25 hidden nodes versus 100 hidden nodes, 25 hidden nodes were selected for use with VATA.

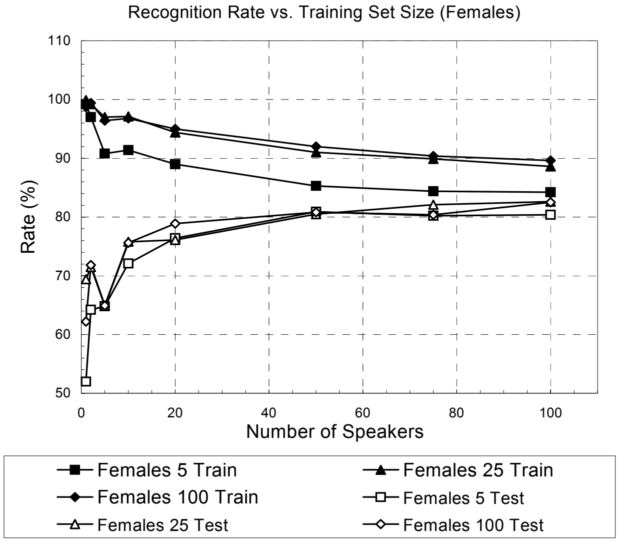

Figure 6. Recognition rate vs. training set size for neural networks with 5, 25 and 100 nodes in hidden the layer (female speakers

The figure above indicates the minimum amount of data that should be used to train a vowel classifer under the conditions used in VATA. In practice, the classifier should be trained using as much data as is available. Table 1 presents vowel classfication results for training data, using all available data for adult male, adult female, child, and general (data from the three other groups combined) speakers. The numbers given indicate an upper limit of expected accuracy of the VATA system used in real time.

Table 1: Training recognition rates for vowels

Speaker Group (speakers in group)	Recognition Rate (%)
Males (114)	89.99
Females (138)	87.33
Children (62)	83.93
General (314)	84.56

5. CONCLUSIONS

Some refinements to the Vowel Articulation Training Aid (VATA), under development at Old Dominion University, have been described. An additional classifier block provides better rejection of out of category sounds than a neural network alone. Experimental data gives an indication of expected accuracy of VATA for real time operation. Interested readers may obtain a copy of the VATA run time program by emailing the first author of this paper (szahoria@odu.edu)

6. REFERENCES

[1] Zahorian S., and Jagharghi, A., (1993) “Spectral-shape features versus formants as acoustic correlates for vowels”, J. Acoust. Soc. Amer. Vol.94, No.4, pp. 1966-1982.

[2] Zahorian S., and Nossair, Z B., (1999) “A Partitioned neural network approach for vowel classification using smoothed time/frequency features”, IEEE Trans. on Speech and Audio Processing, vol. 7, no. 4, pp. 414-425.

[3] Zimmer A., Dai, B., and Zahorian, S, (1998) “Personal Computer Software Vowel Training Aid for the Hearing Impaired”, International Conference on Acoustics, Speech, and Signal Processing, Vol 6, pp. 3625-3628

7. ACKNOWLEDGEMENT

This work was partially supported by NSF grant BES-9977260.